Brown & Bonsaksen, *Cogent Education* (2019), 6: 1571146
https://doi.org/10.1080/2331186X.2019.1571146

cogent education

EDUCATIONAL ASSESSMENT & EVALUATION | RESEARCH ARTICLE

An examination of the structural validity of the Physical Self-Description Questionnaire-Short Form (PSDQ–S) using the Rasch Measurement Model

Ted Brown[1] and Tore Bonsaksen[2,3]*

Abstract: It is important for self-report scales and measures used by educators and health care professionals (such as the Physical Self-Description Questionnaire-Short Form [PSDQ–S]) to have documented reliability and validity. The aim of this study is to investigate the structural validity of the full PSDQ–S composite scale and 11 subscales using the Rasch Measurement Model. 117 healthy children (65 males and 52 females; M = 10 years, 2 months, SD = 1 year, 4 months) completed the PSDQ–S. The PSDQ–S's rating scale functioning, dimensionality, hierarchical ordering, differential item functioning (DIF), and item and person separation reliability were examined. Results supported the scale functioning, dimensionality, hierarchical ordering, DIF, and reliability of the PSDQ–S composite scale and each of its 11 subscales. Therefore, the PSDQ–S composite scale and 11 subscales can be used with confidence to assess children's self-reported physical self-concept.

Subjects: Testing, Measurement and Assessment; Psychometrics/ Testing & Measurement Theory; Test Development, Validity & Scaling Methods; Primary/Elementary Education; Childhood

Keywords: validity; physical self-concept; children; self-report assessment; Rasch Measurement Model

Received: 18 September 2018
Accepted: 13 January 2019
First Published: 23 January 2019

*Corresponding author: Tore Bonsaksen, Department of Occupational Therapy, Prosthetics and Orthotics, Faculty of Health Sciences, OsloMet - Oslo Metropolitan University, PO Box 4 St. Olavs Plass, Oslo 0130, Norway
E-mail: tore.bonsaksen@oslomet.no

Reviewing editor:
Sammy King Fai Hui, Curriculum & Instruction, The Education University of Hong Kong, Hong Kong

Additional information is available at the end of the article

ABOUT THE AUTHORS

Dr Ted Brown is an associate professor and undergraduate course coordinator in the Monash University Department of Occupational Therapy, Frankston, Victoria, Australia. His research interests include occupational therapy practice with children and families, the validation of assessment tools, journal publication bliometrics, health education and evidence-based practice.

Tore Bonsaksen is a professor of occupational therapy at Oslo Metropolitan University in Oslo, Norway, and at VID Specialized University in Sandnes, Norway. His research interests include practice and research related to mental health, public health and health education.

PUBLIC INTEREST STATEMENT

Occupational therapists and educators assess the daily activities of children and adults. It is important that the tests used by therapists accurately and consistently assess the skills and abilities that they claim to. The Physical Self-Description Questionnaire-Short Form (PSDQ–S) is a self-report scale that provides therapists and educators with valuable information about how children perceive themselves. Therefore, investigating the accuracy of the PSDQ–S items is important. Using the PSDQ–S item responses from a sample of 117 healthy children, the quality, accuracy and consistency of the PSDQ–S subscales were investigated. Overall, the PSDQ–S is recommended for use by practitioners, researchers, administrators, and educators.

1. Introduction

All tests, scales, and instruments that are used for clinical, educational or research purposes must exhibit key measurement properties (such as reliability/precision, validity, utility, fairness, clinical utility and responsiveness to change) so that they can be used with confidence, particularly those scales that are high stakes (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME], 2014; Streiner, Norman, & Cairney, 2015). The validity of a test or scale can take many forms and is a continuous, dynamic, and mutual exchange of verification data. "The process of validation involves gathering, synthesising and evaluating data that typically includes confirming and disconfirming evidence" (Cizek, 2016, p. 213). Therefore, the body of evidence of a scale's validity is never static; it is a cumulative, active process with contributions made by the authors of the scale plus other researchers in the cognate field of interest. The Physical Self-Description Questionnaire (PSDQ) (Marsh, Richards, Johnson, Roche, & Tremayne, 1994) and its abbreviated form, referred to as the Physical Self-Description Questionnaire–Short Form (PSDQ–S), are one such set of scales where their validity evidence is dynamic and cyclical. The purpose of this study is to investigate the structural validity of the PSDQ-S using the Rasch Measurement Model (RMM).

2. Literature review

2.1. Physical self-concept

Physical self-concept (also referred to as physical self-perception) refers to a person's perception of themselves in relation to his/her physical skills and outwards appearance (Crocker, Sabiston, Kowalski, McDonough, & Kowalski, 2006). It is divided into perceived physical ability and perceived physical appearance (Babic et al., 2014). "The hierarchical structure of physical self-concept suggests a top-to-bottom hierarchy, where global self-concept is at the apex and actual behavior is at the bottom" (Simons, Capio, Adriaenssens, Delbroek, & Vandenbussche, 2012, p. 874). It has been found that physical self-concept has direct links with many physical and mental health-related issues in children and adolescents; therefore, it is important to have valid and reliable measures of this construct that practitioners, educators and researchers can use (Crocker et al., 2006; Ekeland, Heian, & Hagen, 2005; Utesch, Dreiskämper, Naul, & Geukes, 2018).

2.2. Physical Self-Description Questionnaire (PSDQ)

The PSDQ (Marsh et al., 1994) is a 70-item questionnaire that measures perceived physical self-concept. It was developed in Australia and originally designed for use with adolescents; however, it is also reported to be suitable for children as young as six and seven years of age (Cremeens, Eiser, & Blades, 2006; Klomsten, Skaalvik, & Espnes, 2004). Physical self-concept describes the way that people view themselves physically in areas such as appearance, health, and competence. The PSDQ assesses 11 elements (or factors) of perceived physical self-concept including Strength, Body Fat, Physical Activity, Endurance/Fitness, Sport Competence, Coordination, Health, Appearance, Flexibility, General Physical Self-Concept, and Self-Esteem (Marsh, 1999). The Health and Self-Esteem factors are comprised of eight items each, while all other factors comprise six items.

The PSDQ can be administered individually or with a group of children and takes approximately 20 minutes for children to complete (Marsh et al., 1994). The child is asked to consider each of the 70 statements and circle the number that best describes him/her for each statement. Scoring is performed using the factor structure guide where the PSDQ items generate 11 factor scores. Higher scores indicate greater levels of self-concept.

Evidence of the reliability and validity of the PSDQ has been investigated by Marsh and colleagues and is reported in the empirical literature (Marsh, 1996a, 1997, 1999; Marsh et al., 1994). Their research demonstrates that the PSDQ displays good construct validity (Dishman et al., 2006; Marsh, 1996a; Marsh, Tomás Marco, & Hülya Aþçý, 2002), good test-retest reliability over both

short- and long-term periods (Marsh, 1996b), and good convergent and divergent validity when compared with other self-concept measures (Marsh, 1996a, 1997, 1999; Marsh et al., 1994).

Marsh, Martin, and Jackson (2010) developed a short version of the PSDQ referred to as the PSDQ–S. The PSDQ–S includes 40/70 of the original PSDQ items. Using a cross-validation approach, the 11 PSDQ–S factors (nine distinct subscales and two global measures) exhibited high reliability indices and invariant factor structures across six different participant groups from Australia, Spain, and Israel. Evidence of the PSDQ–S factor validity, test-retest reliability (with correlations ranging from .57 to .90), and convergent and discriminant validity with two other measures of the same dimension, those being the Physical Self Perception Profile and Physical Self Concept instrument, has been reported (Marsh et al., 2010).

The PSDQ–S has been used in several studies (Brewer & Olson, 2015; Martin & Whalen, 2012; Ulrika, Johan, & Guy, 2017) and has been translated into a number of other languages including Chinese, French, Finnish, and Slovenian (Dolenc, 2016; Haapea, Haverinen, Honkalampi, Kuittinen, & Räty, 2016; Maïano, Morin, & Mascret, 2015; Wang, Sun, Liu, Yao, & Pyun, 2015). The PSDQ-2's psychometric properties have only been examined using Classical Test Theory approaches to date (e.g., principal components analysis, traditional confirmatory factor analysis approaches, convergent validity, discriminant validity), but has not been examined through an Item Response Theory lens (such as the Rasch Measurement Model), as of yet.

2.3. Validity

A number of different types of validity have been identified including concurrent validity, predictive validity, content validity, construct validity, criterion validity, and convergent validity (Brown, 2010; Mokkink et al., 2006; Newton & Baird, 2016). In the past, validity was conceptualized as three separate types: content, criterion, and construct with criterion-related validity being subdivided into concurrent and predictive validity (Nunnally & Bernstein, 1994). With the publication of the *Standards of Educational and Psychological Measurement* (*Standards*) (AERA, APA, & NCME) in 1999, validity was redefined following the work of Messick (1989, 1995). Within this framework, it was viewed as a unitary concept, that of construct validity. According to the *Standards* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999) validity referred "to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9) and it discussed five distinct sources of validity evidence: content, response processes, internal structure, relationship to other variables, and consequences of testing. "There are three basic aspects of internal structure: dimensionality, measurement invariance, and reliability" (Rios & Wells, 2014, p. 108). The internal structure source of validity evidence can include factor analytical studies, differential item functioning studies, and item analyses to examine item relationships (AERA, APA, & NCME, 2014; Downing, 2003; Goodwin, 2002).

The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) proposal specified that evidence of six types of measurement properties should be reported about health-related instruments: internal consistency, reliability, measurement error, content validity, construct validity, and criterion validity (Mokkink et al., 2006). More specifically, the components of construct validity evidence with regard to health and epidemiological scales that should be reported according to the COSMIN are structural validity, hypothesis testing, and cross-cultural validity (Mokkink et al., 2006). Structural validity is defined as the degree to which scores of a scale are an adequate indication of the dimensionality of the construct, attribute or factor being measured (Mokkink et al., 2010; Rios & Wells, 2014). Evidence of the structural validity of a scale can take the form of differential item functioning studies and dimensionality studies (Newton & Shaw, 2016). One type of statistical model that can be used to evaluate the structural validity and internal structural validity of an instrument is the Rasch Measurement Model (Bond & Fox, 2015).

2.4. Rasch Measurement Model

The Rasch Measurement Model (RMM), a type of Item Response Theory, is a mathematical model which does not assume that each item of a scale has the same value or replicates the same level of difficulty. The RMM analysis output generates a hierarchical scale of items (easier to hardest) identifying both *person ability* and *item difficulty* (Bond & Fox, 2007). This is based on the principle that how a person responds to an item results from the interaction between the individual and the level of item difficulty. In other words, respondents with higher levels of person ability are expected, based on the premise of the RMM, to answer a larger number of scale items correctly whereas for items with higher levels of difficulty, fewer respondents are expected to answer these items correctly (Lim, Rodger, & Brown, 2009). The RMM provides fit statistics about which scale items fit the Rasch model expectations in order to establish the relationships between the items and the weight of these within the overall construct, and to ascertain if participants responded in a consistent and logical manner. Additionally, where an item is affirmed by fewer respondents, this suggests that a construct may be more challenging to achieve.

Principles of the RMM include *dimensionality, hierarchical ordering, differential item functioning* (DIF), *item and person separation reliability*, and *rating scale functioning* (Bond & Fox, 2015). *Dimensionality* is about examining the individual attributes and how well these fit the suggested model to ascertain whether they do indeed measure the overall construct. This is done within a hierarchical line of enquiry of *more than/less than* and is confirmed when the statistics fit within the acceptable range and if most of the variance can be explained by the data thereby indicating unidimensionality (Boone, Staver, & Yale, 2014).

Fit statistics (*infit* and *outfit*) of the data to the model can be determined if each item matches the expectations of the RMM. Statistical values that are close to 1.0 are considered a good fit. These are reported in the form of mean square scores (MNSQ ranges 0.6 to 1.4) and standardised mean scores (ZSTD range −2 to +2) (Bond & Fox, 2015). Usually, misfitting items are identified and removed from the model. The goal is to keep only the scale items that function substantially in measuring the underlying scale construct.

Hierarchical ordering of item difficulty is also explored through the application of the RMM to scale items (Lim et al., 2009). The results identify the hierarchy in which people find the items least to most difficult to answer correctly. The fit statistics determine the logit value (log-odds units) to provide a consistent value and meaning of the intervals within the scale (Bond & Fox, 2015). This has relevance when there are many items within a test, since having a range of items with a range of challenge is what is sought.

Differential item functioning (DIF) is about identifying if the constructs establish consistent difficulty of the items when used with different groups of people to determine if the test items are biased (Bond & Fox, 2015). In other words, with the RMM, DIF is investigated by examining whether individual scale items function differently across various subgroups (de Ayala, 2009). For example, DIF can be explored across a number of respondent-related variables including gender, age, ethnicity, socioeconomic status, level of educational achievement, region of residence, or diagnosis. If participants respond differently across the items, then an inherent bias may be present (Boone et al., 2014). It is advisable for clinicians not to use scales with known DIF towards participant groups exhibiting specific traits since the scale results may be biased, skewed or not valid (e.g., scales used with ethnic minority groups or scales relevant to one cultural/age group and not another). DIF has also been referred to as parameter invariance (Bond & Fox, 2015).

RMM statistics programs provide *item and person separation reliability*. Low person separation infers that a scale may not be sensitive enough to differentiate between persons with low and high ability. Item separation is used to examine the item hierarchy. Low item separation suggests that the sample of persons may not be large enough to demonstrate the item difficulty hierarchy. *Rating response scale functioning* refers to how well a rating scale that participants answer

represents a continuum of the construct being measured. In other words, at one end of the response scale participants have less of the construct and at the other end of the scale participants have more of the construct (Boone et al., 2014).

3. Purpose

The purpose of this study is to examine the *dimensionality, hierarchical ordering, differential item functioning* (DIF), *item and person separation reliability*, and *rating scale functioning* of the PSDQ–S's 11 subscales using the RMM methodology with a sample of school-age children. It is hypothesized that (1) the dimensionality of the PSDQ–S composite scale and 11 subscales will be confirmed; (2) there will be a lack of DIF of the item calibrations of the PSDQ–S composite scale and 11 subscales across different groups of participants based on gender (e.g., males versus females); (3) the items of the PSDQ–S composite scale, two global scales and nine factors will form hierarchical indices; and (4) the PSDQ–S composite scale and 11 subscales will exhibit adequate levels of item and person separation reliability. This in turn will provide evidence about the *structural validity* and reliability of the PSDQ–S composite scale and 11 subscales.

4. Method

4.1. Design

A quantitative cross sectional scale validation design using the Rasch Measurement Model was used to complete this study.

4.2. Participants

A convenience sample of 117 children were recruited for this study from Geelong and Melbourne metropolitan regions, Victoria, Australia. Inclusion criteria were that the children were between 8 and 12 years of age, had a working knowledge of the English language, and had consent provided by a parent/guardian for participation in the study. Children themselves also had to provide verbal assent to take part in the study. Children were excluded if they presented with a known history of physical, psychosocial or intellectual impairment (based on parental report) that would impact upon their ability to complete the PSDQ–S.

4.3. Instrumentation

The Physical Self-Description Questionnaire–Short Form (PSDQ–S) was developed by Marsh et al. (2010) to measure the multiple dimensions of physical self-concept and is composed of 40 items that measure nine discrete sub-factors and two global factors: Strength, Body Fat, Physical Activity, Endurance/Fitness, Sport Competence, Coordination, Health, Appearance, Flexibility, General Physical Self-Concept, and Self-Esteem. Each PSDQ–S item is scored on a six-point Likert scale anchored by 1: False and 6: True. Respondents are asked to "Please circle the number which is the most correct statement about you" (Marsh et al., 2010, p. 481). Some of the PSDQ–S items are negatively phrased as a way to minimize respondents from answering the items in a socially desirable manner and these are reverse scored. The responses that relate to each global factor and sub-factor are averaged; therefore, the maximum average score for each sub-factor that can be achieved is six (Marsh et al., 2010).

Details of the PSDQ–S's reliability and validity have been reported in the literature (Dolenc, 2016; Haapea et al., 2016; Maïano et al., 2015; Marsh et al., 2010; Wang et al., 2015). Evidence of its internal consistency (α range 81–. 94, median α = 0.92), and test stability/test-retest reliability (r = 0.69, 14 months; r = 0.83, 3 months) with populations from several different cultural contexts (e.g., Australia, Spain, and Israel) and respondent age groups have been reported (Marsh et al., 2010). Verification of the PSDQ–S's construct, convergent, content, and discriminant validity have also been reported in the peer-reviewed literature (Maïano et al., 2015; Marsh et al., 2010).

4.4. Data analysis

The *Statistical Package for the Social Sciences Version 22* (SPSS) was used for the data entry, its storage, and retrieval. Descriptive statistics such as measures of central tendency and measures of variance were calculated as appropriate to the data using SPSS. The RMM computer program, *Winsteps (version 3.70.0)* (Linacre, 2011) was used for the data analysis (Linacre & Wright, 1998). RMM analysis is an iterative process with the objective of achieving "best fit" of the data to the model by testing the model's assumptions. The intent of the RMM analysis were to determine (1) the dimensionality of the PSDQ–S's composite scale and 11 subscales based on goodness-of-fit analysis; (2) whether differential item functioning (DIF) of the item calibration estimates occurred across participant samples in terms of gender for any of the items of the PSDQ–S's composite scale and 11 subscales; (3) the hierarchical ordering and spacing of the PSDQ–S's composite scale and 11 subscales based on item calibration (item difficulty parameter estimate); and (4) the item and person separation reliability of the PSDQ–S's composite scale and 11 subscales.

4.5. Rasch Measurement Model analysis procedures

4.5.1. Item fit

The RMM evaluates the fit of the data to an unconditional probabilistic model. The logit values represent the difficulty of the items (item weights) in an instrument and items are ordered from easiest to most difficult, providing evidence of hierarchical ordering of scale items. The fit of the items to the RMM was determined by the infit mean square statistic (MNSQ) and the outfit mean square statistic, both of which are based on a chi-square distribution (Smith, 1992). Fit statistics should range between 0.60 and 1.40 to fit RMM expectations (Linacre, 2013). High or low fit statistics represent abnormalities in the response pattern to the item that may be related to a lack of dimensionality, DIF, poorly placed items in terms of developmental sequencing, or poorly worded items (Linacre & Wright, 1998). This step indicates how items fit the RMM. Item fit is also one step in the confirmation of dimensionality.

The infit and outfit statistics use slightly different methods for assessing an item's fit to the RMM. The infit statistic gives more weight to the performance scores of subjects closer to the item value. The belief is that persons whose ability is close to the item's difficulty will give a more sensitive insight into that item's performance (Boone et al., 2014). The outfit statistic is not weighted, and therefore is more sensitive to the influence of outlying scores.

4.5.2. Confirmation of dimensionality

Dimensionality, which will be partially established through item fit, will be further confirmed by principal component analysis with orthogonal Varimax rotation of the item residuals. Factor analysis is a mathematical process that determines linear combinations of the variables in order to explain the maximum amount of variance in the data (Nunnally & Bernstein, 1994; Yong & Pearce, 2013). The first factor to explain the largest amount of variance is represented by the construct identified by the RMM analysis. Hence, factor analysis of the item residuals should not identify any additional factors (e.g., minimal component of variance explained) if the assumption of dimensionality is upheld. The criterion specified for the percentage of the raw variance a factor must account for in order for the items in that factor to be considered to compromise a unidimensional measure was 60% (Bond & Fox, 2015; Thompson, 2004) with the secondary dimension accounting for less than 5% (Smith & Miao, 1994). The criterion specified for the minimum factor loading an item can have and still be considered part of the underlying latent trait was 0.40 (Nunnally & Bernstein, 1994).

4.5.3. Item hierarchical ordering

The average item calibrations from the RMM analysis (referred to as logits) defined the hierarchical order of the items along the continuum (Boone et al., 2014). The items of the PSDQ–S composite scale and 11 subscales are mapped onto an item difficulty map based on their respective logit scores. Harder items were located at one end of the linear continuum and easier items were

located at the opposite end. This provides evidence of the hierarchical ordering of the PSDQ–S composite scale and 11 subscale items.

4.5.4. Differential Item Functioning (DIF)

DIF, as evaluated by examining the difference between logit scores for each of the PSDQ–S composite scale and 11 subscale items based on gender. In this instance, logit values of the scale items based first on gender (males versus females) were generated and examined for potential significant differences using t-test comparisons. If any significant differences were found between the two sets of RMM logit scores, then DIF would be present (Bond & Fox, 2015).

4.5.5. Reliability

The *Winsteps* program generates both person reliability statistics and item reliability statistics. Person reliability depends mainly on sample ability variance, length of the scale, number of rating categories per item, and sample-item targeting. Item reliability is primarily dependent on item difficulty variance and participant sample size. The item and person separation indexes were converted into strata according to the formula [4(separation index) +1]/3 (Wright & Masters, 1982). The strata were utilized to pinpoint the number of discreet groups of items (based on difficulty) and people (based on ability). It is expected that scales should separate the items and people into at least two separate groups. The item reliability coefficient should ideally be >0.80 and the item separation index (ISI) should be >3.0. The person reliability coefficient should ideally be >0.80, the person separation index (PSI) should be >2.0, and the person raw score reliability should be >0.80 (Arnadottir & Fisher, 2008; Wright & Masters, 1982).

4.6. Procedures

Ethical approval for this study was granted by Monash University Human Research Ethics Committee and the Victorian Department of Education Human Research Ethics Committee. Four state primary state schools within the Geelong and Melbourne metropolitan regions were approached to participate in the study. The primary schools were chosen for their differing metropolitan locations to increase the diversity of the sample. Consent was obtained from principals at three of the schools. After consent was obtained, information packages and consent forms were provided to school administration staff who then randomly distributed the packages to 250 students between eight and 12 years of age. Parents were asked to return signed consent forms in a reply-paid envelope to indicate willingness for both the parent and the child to participate in the study. A total of 156 signed consent forms were returned to the researcher; however, 39/156 children were excluded due to not meeting the inclusion criteria for the study. The final sample size for the study was 117 participants.

During one session the researcher met with each child individually for approximately 20 minutes to complete the PSDQ–S. Prior to each session the researcher explained the purpose of the session and sought the child's verbal consent to participate. All 117 children provided verbal consent to participate in the study. The data collection sessions were completed within school grounds at a time negotiated with the child's classroom teacher to ensure minimal impact upon the child's learning. The majority of sessions were conducted indoors within a spare classroom.

5. Results

5.1. Participants

The sample comprised 117 children (a response rate of 46.8%) of whom 65 were males (55.6%) and 52 females (44.4%). Participants varied in age from eight years to 12 years, 2 months, with a mean age of 10 years, 2 months (standard deviation [SD] = 1 year, 4 months).

5.2. RMM item fit

The PSDQ–S composite scale and 11 subscales mean scores, SDs, and interquartile range are located in Table 1. The RMM fit of the PSDQ–S composite scale and 11 subscale items were

Table 1. Physical Self-Description Questionnaire-Short Form (PSDQ–S) subscale descriptive statistics (N = 117)

	PA	AP	BF	CO	EN	SE	FL	GP	HE	SP	ST	Total Scale Score
Mean	19.11	12.28	11.41	24.58	13.99	22.63	12.72	15.60	18.32	15.16	14.17	179.98
Std. Deviation	4.40	3.18	6.25	4.44	3.37	4.87	3.80	2.55	8.58	3.15	3.14	27.07
Range	15	15	15	16	12	20	15	11	25	15	12	129
25th Percentile	16	10	3	22	12	18	10	14	10	13	12	164
50th Percentile	20	12	13	26	14	22	13	16	19	16	15	177
75th Percentile	23	15	18	28	17	27	15	18	27	18	17	198

PA = Activity; AP = Appearance; BF = Body Fat; CO = Coordination; EN = Endurance; SE = Global Esteem;
FL = Flexibility; GP = General Physical; HE = Health; SP = Sport; ST = Strength

Table 2. Total Physical Self-Description Questionnaire-Short Form (PSDQ–S) Rasch Measurement Model (RMM) composite item statistics (N = 117)

PSDQ–S Scale Items	RMM Logit Item Measure	Logit Item Measure S. E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Point Measure Correlation
PSDQ–S 34	.88	.07	1.37	2.6	1.24	1.5	.56
PSDQ–S 31*	**.75**	**.07**	**1.49**	**3.7**	**1.41**	**2.7**	**.52**
PSDQ–S 32	.73	.07	1.33	2.6	1.25	1.7	.54
PSDQ–S 20	.73	.07	1.30	2.4	1.23	1.6	.55
PSDQ–S 10	.68	.06	1.42	3.3	1.33	2.2	.53
PSDQ–S 23	.67	.06	1.32	2.6	1.27	1.9	.51
PSDQ–S 8	.66	.06	1.33	2.7	1.27	1.9	.52
PSDQ–S 33	.62	.06	.98	−.1	.96	−.3	.49
PSDQ–S 9	.60	.06	1.38	3.2	1.33	2.4	.48
PSDQ–S 6	.51	.06	.79	−2.1	1.30	2.2	.39
PSDQ–S 30	.49	.06	1.12	1.1	1.32	2.3	.36
PSDQ–S 26	.11	.07	.80	−1.9	.99	.0	.37
PSDQ–S 25	.10	.07	1.06	.5	1.74	4.2	.26
PSDQ–S 7	−.02	.07	.74	−2.3	.83	−1.1	.37
PSDQ–S 5	−.02	.07	.89	−.9	1.48	2.7	.27
PSDQ–S 1	−.03	.07	.99	.0	1.19	1.2	.35
PSDQ–S 17	−.03	.07	.81	−1.6	.89	−.7	.41
PSDQ–S 24	−.03	.07	.82	−1.5	1.08	.6	.36
PSDQ–S 40	−.06	.07	.85	−1.2	.95	−.2	.35
PSDQ–S 16	−.10	.07	.84	−1.2	1.07	.4	.36
PSDQ–S 38	−.14	.07	.79	−1.7	.93	−.4	.32
PSDQ–S 4	−.19	.07	1.01	.1	1.11	.7	.37
PSDQ–S 19	−.20	.07	.64	−2.9	.67	−2.0	.37
PSDQ–S 14	−.26	.08	.62	−2.9	.68	−1.9	.42
PSDQ–S 13	−.29	.08	.70	−2.2	.73	−1.5	.40
PSDQ–S 18	−.29	.08	.73	−1.9	.89	−.6	.37
PSDQ–S 15	−.29	.08	.78	−1.5	.74	−1.5	.38
PSDQ–S 3	−.30	.08	1.22	1.4	1.22	1.2	.35
PSDQ–S 35	−.31	.08	.89	−.7	1.04	.3	.36
PSDQ–S 39	−.33	.08	.86	−.9	1.05	.3	.37
PSDQ–S 11	−.34	.08	.70	−2.1	.77	−1.2	.37
PSDQ–S 12	−.35	.08	1.00	.1	1.11	.6	.30
PSDQ–S 36	−.36	.08	.86	−.9	1.09	.5	.35
PSDQ–S 2	−.38	.08	1.16	1.0	1.33	1.6	.32
PSDQ–S 21	−.42	.08	.64	−2.4	.70	−1.6	.36
PSDQ–S 27	−.45	.09	.93	−.4	.93	−.3	.33
PSDQ–S 28	−.53	.09	.78	−1.3	.89	−.5	.33
PSDQ–S 37	−.53	.09	.67	−2.0	.63	−1.9	.41
PSDQ–S 29	−.59	.09	.85	−.8	1.02	.2	.30
PSDQ–S 22	−.73	.10	.77	−1.1	.80	−.8	.30

* misfitting item according to RMM requirements of MNSQ range between 0.60–1.4, ZSDT range between −2 and 2, and/or Point Measure Correlation <0.20

Note: RMM = Rasch Measurement Model; PSDQ–S = Physical Self-Description Questionnaire-Short Form; MNSQ = Mean Square; ZSTD = z-standardized

Table 3. Physical Self-Description Questionnaire-Short Form (PSDQ–S) Rasch Measurement Model (RMM) subscale item statistics (N = 117)

PSDQ–S Scale Items	RMM Logit Item Measure	Logit Item Measure S. E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Point Measure
Correlation							
Activity (PA)							
PSDQ–S 1	.35	.09	1.40	3.2	1.32	1.9	.70
PSDQ–S 4	.06	.10	.62	−2.8	.63	−2.5	.76
PSDQ–S 3	−.14	.10	.89	−.7	.82	−1.1	.70
PSDQ–S 2	−.28	.10	1.21	1.3	1.13	.8	.63
Appearance (AP)							
PSDQ–S 6	1.41	.13	1.21	1.5	1.23	1.6	.79
PSDQ–S 7	−.70	.13	.73	−2.2	.70	−2.3	.85
PSDQ–S 5	−.71	.13	1.06	.5	1.01	.1	.81
Body Fat (BF)							
PSDQ–S 8	.12	.16	.95	−.2	.88	−.5	.86
PSDQ–S 9	−.06	.16	1.1	.1	1.01	.1	.85
PSDQ–S 10	−.06	.16	.98	0	.92	−.3	.33
Coordination (CO)							
PSDQ–S 14	.11	.12	.84	−1.0	.86	−.9	.72
PSDQ–S 13	.04	.12	.80	−1.4	.88	−.8	.74
PSDQ–S 15	.03	.12	.71	−2.1	.63	−2.7	.77
PSDQ–S 11	−.07	.12	1.22	1.4	1.28	1.7	.62
PSDQ–S 12	−.10	.12	1.41	2.4	1.38	2.2	.65
Endurance (EN)							
PSDQ–S 17	.35	.12	.94	−.3	.93	−.5	.84
PSDQ–S 16	.11	.13	1.06	.4	1.01	.1	.81
PSDQ–S 18	−.46	.13	.95	−.3	.95	−.3	.78
Global Esteem (SE)							
PSDQ–S 20	.89	.08	.69	−2.3	.63	−2.4	.80
PSDQ–S 23	.81	.08	.95	−.3	.99	.0	.75
PSDQ–S 19	−.25	.08	.68	−2.3	1.21	1.2	.44
PSDQ–S 21	−.52	.09	.68	−1.8	1.03	.2	.42
PSDQ–S 22	−.93	.12	1.19	.9	1.30	1.3	.31
Flexibility (FL)							
PSDQ–S 26	.19	.13	1.01	.1	.99	.0	.85
PSDQ–S 25	.16	.13	.98	−.1	.90	−.7	.86
PSDQ–S 24	−.34	.13	.98	−.1	.97	−.2	.84
General Physical (GP)							
PSDQ–S 27	.19	.14	1.06	.4	1.03	.3	.78
PSDQ–S 28	−.02	.15	.91	−.4	.95	−.2	.78
PSDQ–S 29	−.17	.15	.98	.0	1.04	.3	.77
Health (HE)							
PSDQ–S 34	.41	.10	1.03	.2	.75	−1.2	.83
PSDQ–S 31	.11	.10	1.09	.6	.81	−1.0	.82

(*Continued*)

PSDQ–S Scale Items	RMM Logit Item Measure	Logit Item Measure S. E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Point Measure
PSDQ–S 32	.08	.10	.71	−1.9	.56	−2.6	.86
PSDQ–S 33	−.17	.09	.88	−.8	.91	−.4	.85
PSDQ–S 30	−.44	.09	1.33	2.1	1.83	3.7	.76
Sport (SP)							
PSDQ–S 35	.43	.17	.85	−.8	.85	−.9	.90
PSDQ–S 36	.15	.18	1.09	.5	1.08	.5	.86
PSDQ–S 37	−.58	.19	.96	−.2	.83	−1.0	.86
Strength (ST)							
PSDQ–S 40	.37	.13	1.04	.3	1.01	.1	.83
PSDQ–S 38	.12	.13	.78	−1.6	.80	−1.4	.84
PSDQ–S 39	−.49	.14	1.24	1.5	1.10	.7	.77

* misfitting item according to RMM requirements of MNSQ range between 0.60–1.4 and ZSDT range between −2 and 2
Note: RMM = Rasch Measurement Model; PSDQ–S = Physical Self-Description Questionnaire-Short Form; MNSQ = Mean Square; ZSTD = z-standardized; PA = Activity; AP = Appearance; BF = Body Fat; CO = Coordination; EN = Endurance; GE = Global Esteem; FL = Flexibility; GP = General Physical; HE = Health; SP = Sport; ST = Strength

assessed using fit mean squares statistics and standardized fit statistics (ZSDT) (see Table 2). Mean square (MNSQ) values outside the range of 0.6 to 1.4 and ZSTD scores outside the +2 to −2 range were identified as potential misfitting items.

The item logit scores for the PSDQ–S composite scale ranged from −0.73 to 0.88 (see Table 2) while for the 11 subscales the logit scores ranged from −0.93 to 1.41 (see Table 3). The PSDQ–S Coordination subscale had the narrowest item logit score range (0.21 logits), whereas the PSDQ–S Global Esteem subscale had the largest item logit score range (1.82 logits) (see Table 3). For the PSDQ–S composite scale, one item was found to exhibit RMM misfit; PSDQ–S Item 31 had an Infit MNSQ score of 1.49 (Infit ZSTD 3.7) and an Outfit MNSQ score of 1.41 (Outfit ZSTD 2.7) (see Table 2). All the items for the 11 PSDQ–S subscales met the RMM fit requirements with Infit MNSQ scores ranging from 0.62 to 1.14, Infit ZSTD scores ranging from −2.8 to 3.2, Outfit MNSQ scores ranging from 0.63 to 1.83, and Outfit ZSTD scores ranging from −2.7 to 3.7 (see Table 3). Point measure correlations for the composite PSDQ–S items ranged from 0.26 to 0.56 (see Table 2), whereas for the 11 PSDQ–S subscales, the point measure item correlations ranged from 0.31 to 0.90 (see Table 3).

For the PSDQ–S composite scale, the item separation index was 5.71, item spread was 1.61, the number of separate item strata was 8, and the item mean was 0.0 logits (SD = 0.45) (see Table 4). For the PSDQ–S composite scale, the person separation index was 3.05, person spread was 3.82, the number of separate person strata was 4.40, and the person mean was 0.46 (SD = 0.60) (see Table 4).

For the 11 PSDQ–S subscales, the item separation indices ranged from 1.08 to 7.58, item spread ranged from 0.18 to 2.12, the number of separate item strata ranged from 1.8 to 10.1, and the item mean ranged from 0.00 (SD = 0.00) to 0.0 logits (SD = 0.73) (see Table 4). For the 11 PSDQ–S subscales, the person separation indices ranged from 0.95 to 1.86, person spread ranged from 3.22 to 7.69, the number of separate person strata ranged from 1.7 to 3.9, and the person mean ranged from 0.41 (SD = 0.80) to 1.73 (SD = 2.00) (see Table 4).

Table 4. Parameters of the Physical Self-Description Questionnaire-Short Form (PSDQ–S) composite and subscales with Rasch Measurement Model (RMM) requirements (N = 117)

Parameter	RMM requirements	PSDQ–S total scale (40 items)	PSDQ–S Activity subscale (4 items)	PSDQ–S Appearance subscale (3 items)	PSDQ–S Body Fat subscale (3 items)	PSDQ–S Coordination subscale (5 items)	PSDQ–S Endurance subscale (3 items)
Model Requirements							
• Monotonicity	Visual inspection of item thresholds; all thresholds order	✓	✓	✓	✓	✓	✓
• Local Independence	r >0.3 between the standardized residuals of the Rasch analysis between any item pairs would violate local independence	✓	✓	✓	✓	✓	✓
• Unidimensionality	PCA analysis of item residuals	✓	✓	✓	✓	✓	✓
• Differential Item Functioning	Percentage of items demonstrating DIF is <5%	× one item exhibited DIF	✓	✓	✓	✓	✓
Model Fit: summary of items							
• Item hierarchy: face validity	Item hierarchical ordering conforms to theoretical/ clinical expectations	✓	✓	✓	✓	✓	✓
• Item mean (SD) logits	0.0	0.00 (0.45)	0.00 (0.24)	0.00	0.00 (0.08)	0.00 (0.08)	0.0 (0.34)
• Item reliability*	>0.8	0.97 [excellent]	0.80 [fair]	0.98 [excellent]	0.95 [excellent]	0.94 [very good]	0.86 [very good]
• Item separation index	>3.0	5.71	1.98	7.34	1.28	1.25	2.44
• Item spread		1.61	0.60	2.12	0.18	0.21	0.81

(*Continued*)

cogent education

Table 4. (*Continued*)

Parameter	RMM requirements	PSDQ–S total scale (40 items)	PSDQ–S Activity subscale (4 items)	PSDQ–S Appearance subscale (3 items)	PSDQ–S Body Fat subscale (3 items)	PSDQ–S Coordination subscale (5 items)	PSDQ–S Endurance subscale (3 items)
• Item Model Init MNSQ Range Extremes*	0.60–1.4	0.62–1.49 [fair]	0.62–1.40 [very good]	0.73–1.21 [very good]	0.95–1.01 [fair]	.71–1.41 [very good]	0.94–1.06 [fair]
• Item Model Infit ZSTD Range Extremes	−2.0–+2.0	(−2.9) to 3.7	(−2.8) to 2.5	(−2.2) to 1.5	(−3.8) to 2.9	(−2.1) to 2.4	(−0.3) to 0.5
• Item Model Outfit MNSQ Range Extremes*	0.60–1.4	0.63–1.74 [fair]	0.63–1.32 [good]	0.7–1.23 [very good]	0.5–2.42 [good]	0.63–1.38 [good]	0.93–1.01 [fair]
• Item Model Outfit ZSTD Range Extremes	−2.0–+2.0	(2.0) to 4.2	(−2.5) to 1.9	(−2.3) to 16	(−3.6) to 6.4	(−2.7) to 2.2	(−0.5) to 0.1
• # misfitting items	0	1 (2.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
• PSDQ-S misfitting items		31	NA	NA	NA	NA	NA
• # Separate Item Strata*	>3	8.0 [good]	3.0 [fair]	10.1 [excellent]	2.0 [poor]	2.0 [poor]	3.6 [good]
Model Fit: summary of persons							
• Person spread		3.82	3.22	7.69	6.42	4.2	5.34
• Person mean (SD) logits		0.46 (0.60)	0.74 (0.85)	0.94 (1.75)	0.51 (1.79)	1.47 (1.20)	1.04 (1.39)
• # misfitting persons	<5%	1 (0.85%)	28 (23.9%)	6 (5.1%)	69 (59%)	14 (12%)	26 (22.2%)
Measurement quality: reliability and targeting							
• Person reliability*	>0.80 for individual measurement	0.90 [very good]	0.51 [poor]	0.71 [fair]	0.78 [fair]	0.64 [poor]	0.62 [poor]

(*Continued*)

cogent education

Table 4. (*Continued*)

Parameter	RMM requirements	PSDQ–S total scale (40 items)	PSDQ–S Activity subscale (4 items)	PSDQ–S Appearance subscale (3 items)	PSDQ–S Body Fat subscale (3 items)	PSDQ–S Coordination subscale (5 items)	PSDQ–S Endurance subscale (3 items)
• Person Separation Index (PSI)	>2.0	3.05	1.03	1.56	1.86	1.35	1.27
• # of Separate Person Strata*	>3 strata for individual measurement	4.4 [very good]	1.7 [poor]	2.4 [fair]	2.8 [fair]	2.1 [fair]	2.0 [fair]
• Person Raw Score Reliability	>0.80 for individual measurement	0.91	0.75	0.77	0.96	0.81	0.79
• Standard Error of Measurement (SEM)	Standard Error of Measurement (SEM) as low as possible	0.20	0.60	0.94	0.84	0.72	0.86
• Precision of the person measure estimates	Expected to be higher in the middle of the measurement range	25	2.78	1.17	1.43	1.93	1.35
• Targeting—Targeting Index*	When average person measure is [−1, 1] targeting is good; when [−2, 2] targeting is fair	0.46: 25 [fair]	0.74: 0.60 [good]	0.94: 1.17 [good]	0.51: 0.84 [fair]	1.47: 0.72 [poor]	1.04: 0.86 [fair]
• Ceiling effect (% of persons with maximum score)*	<2–5%	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]
• Floor effect (% of persons with minimum score)*	<2–5%	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]
• Difference between person & item means	<1.0 logit	0.46	0.74	0.94	0.51	1.47	1.04

(*Continued*)

Table 4. (*Continued*)

Parameter	RMM requirements	PSDQ–S total scale (40 items)	PSDQ–S Activity subscale (4 items)	PSDQ–S Appearance subscale (3 items)	PSDQ–S Body Fat subscale (3 items)	PSDQ–S Coordination subscale (5 items)	PSDQ–S Endurance subscale (3 items)
Dimensionality							
• Variance accounted for by 1st factor*	>60%	69.2% [good]	61.5% [good]	78.7% [very good]	84.1% [excellent]	60.1% [good]	67.9%[good]
• PCA (eigenvalue for 1st contrast)	≤2.0	9.6	1.6	1.7	1.6	1.6	1.6
• Unexplained variance in contrasts 1–5 of PCA of residuals*	<5%	1.4% [excellent]	0.0% [excellent]	0.0% [excellent]	0.0% [excellent]	0.0% [excellent]	0.0% [excellent]
Differential Item Functioning							
• DIF by gender (total # of items [% of total number of items])	>0.5 logits; $p < .05$	1 (2.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
• Item # exhibiting DIF by gender		25	NA	NA	NA	NA	NA
Model Requirements							
• Monotonicity	Visual inspection of item thresholds; all thresholds order	✓	✓	✓	✓	✓	✓
• Local Independence	r >0.3 between the standardized residuals of the Rasch analysis between any item pairs would violate local independence	✓	✓	✓	✓	✓	✓

(*Continued*)

cogent education

Table 4. (*Continued*)

Parameter	RMM requirements	PSDQ–S total scale (40 items)	PSDQ–S Activity subscale (4 items)	PSDQ–S Appearance subscale (3 items)	PSDQ–S Body Fat subscale (3 items)	PSDQ–S Coordination subscale (5 items)	PSDQ–S Endurance subscale (3 items)
• Unidimensionality	PCA analysis of item residuals	✓	✓	✓	✓	✓	✓
• Differential Item Functioning	Percentage of items demonstrating DIF is <5%	✓	✓	× one item exhibited DIF	✓	✓	✓
Model Fit: summary of items							
• Item hierarchy: face validity	Item hierarchical ordering conforms to theoretical/ clinical expectations	✓	✓	✓	✓	✓	✓
• Item mean (SD) logits	0.0	0.0 (0.73)	0.0 (0.24)	0.0 (0.15)	0.0 (0.29)	0.0 (0.43)	0.0 (0.36)
• Item reliability*	>0.8	0.98 [excellent]	0.71 [fair]	0.61[poor]	0.88 [good]	0.82 [good]	0.85 [good]
• Item separation index	>3.0	7.58	1.57	1.08	2.67	2.11	2.37
• Item spread		1.82	0.53	.32	0.85	1.01	0.86
• Item Model Infit MNSQ Range Extremes*	0.60–1.4	0.68–1.19 [fair]	0.98–1.01 [fair]	0.91–1.06 [fair]	0.71–1.33 [very good]	0.85–1.09 [fair]	0.78–1.24 [very good]
• Item Model Infit ZSTD Range Extremes	−2.0–+2.0	−2.3–0.9	−0.1–0.1	−0.4–0.4	−1.9–2.1	−0.8–0.5	−1.6–1.5
• Item Model Outfit MNSQ Range Extremes*	0.60–1.4	0.63–1.30 [good]	0.90–0.99 [poor]	0.95–1.04 [poor]	0.56–1.83 [poor]	0.83–1.08 [poor]	0.80–1.10 [poor]
• Item Model Outfit ZSTD Range Extremes	−2.0–+2.0	−2.4–1.3	−0.7 to 0.6	−0.20–0.3	−2.6–3.7	−1.0–0.50	−1.4–0.7
• # misfitting items	0	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

(*Continued*)

cogent education

Table 4. (*Continued*)

Parameter	RMM requirements	PSDQ–S total scale (40 items)	PSDQ–S Activity subscale (4 items)	PSDQ–S Appearance subscale (3 items)	PSDQ–S Body Fat subscale (3 items)	PSDQ–S Coordination subscale (5 items)	PSDQ–S Endurance subscale (3 items)
• PSDQ–S misfitting items		NA	NA	NA	NA	NA	NA
• # Separate Item Strata*	>3	10.4 [excellent]	2.4 [fair]	1.8 [poor]	3.9 [good]	3.2 [good]	3.5 [good]
Model Fit: summary of persons							
• Person spread		3.81	7.76	4.66	4.87	7.12	5.69
• Person mean (SD) logits		0.41 (0.80)	0.76 (1.76)	1.35 (1.37)	0.45 (1.32)	1.73 (2.00)	1.18 (1.47)
• # misfitting persons	<5%	10 (8.5%)	18 (15%)	41 (35%)	11 (9.4%)	45 (38.5%)	25 (21.4%)
Measurement quality: reliability and targeting							
• Person reliability*	>0.80 for individual measurement	0.57 [poor]	0.73 [fair]	0.47 [fair]	0.77 [fair]	0.69 [fair]	0.62 [poor]
• Person Separation Index (PSI)	>2.0	1.16	1.63	0.95	1.85	1.48	1.28
• # of Separate Person Strata*	>3 strata for individual measurement	1.9 [poor]	2.5 [fair]	1.6 [poor]	2.8 [fair]	3.9 [good]	2.0 [fair]
• Person Raw Score Reliability	>0.80 for individual measurement	0.60	0.84	0.72	0.91	0.88	0.78
• Standard Error of Measurement (SEM)	Standard Error of Measurement (SEM) as low as possible	0.34	0.91	0.90	0.63	1.11	0.91

(*Continued*)

cogent education

Table 4. (*Continued*)

Parameter	RMM requirements	PSDQ–S total scale (40 items)	PSDQ–S Activity subscale (4 items)	PSDQ–S Appearance subscale (3 items)	PSDQ–S Body Fat subscale (3 items)	PSDQ–S Coordination subscale (5 items)	PSDQ–S Endurance subscale (3 items)
• Precision of the person measure estimates	Expected to be higher in the middle of the measurement range	8.65	1.21	1.23	2.73	0.81	1.21
• Targeting—Targeting* Index	When average person measure is [−1, 1] targeting is good; when [−2, 2] targeting is fair	0.41: 0.34 [good]	0.76: 0.91 [fair]	1.36: 0.90 [poor]	0.45: 0.63 [fair]	1.73: 1.11 [poor]	1.18: 1.21 [good]
• Ceiling effect (% of persons with maximum score)*	<2–5%	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]
• Floor effect (% of persons with minimum score)*	<2–5%	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]	0% [excellent]
• Difference between person & item means	<1.0 logit	0.41	0.76	1.35	0.45	1.73	1.18
Dimensionality							
• Variance accounted for by 1st factor*	>60%	65.9% [good]	76% [very good]	52.1% [poor]	87.8% [excellent]	66.9% [good]	64.2% [poor]
• PCA (eigenvalue for 1st contrast)	≤2.0	1.4	1.6	1.6	2.3	1.7	1.6
• Unexplained variance in contrasts 1–5 of PCA of residuals*	<5%	0% [excellent]	0% [excellent]	0% [excellent]	0.1% [excellent]	0% [excellent]	0% [excellent]

(*Continued*)

cogent • education

Table 4. (*Continued*)

Parameter	RMM requirements	PSDQ–S total scale (40 items)	PSDQ–S Activity subscale (4 items)	PSDQ–S Appearance subscale (3 items)	PSDQ–S Body Fat subscale (3 items)	PSDQ–S Coordination subscale (5 items)	PSDQ–S Endurance subscale (3 items)
Differential Item Functioning							
• DIF by gender (total # of items [% of total # of items])	>0.5 logits; $p < .05$	0 (0.0%)	0 (0.0%)	1 (33.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
• Item number exhibiting DIF by gender		NA	NA	29	NA	NA	NA

RMM = Rasch Measurement Model; # = number; SD = Standard Deviation; PCA = Principal Components Analysis; DIF = Differential Item Functioning; MNSQ = Mean Square. NA = not applicable. *Person spread* is defined as the difference between maximum person logit score and minimum person logit score. *Item spread* is defined as the difference between maximum item logit score and minimum item logit score. ✓ denotes that the RMM quality assessment guideline requirement has been met, whereas × infers that the quality assessment guideline requirement has not been met.

* refers to Rating Scale Instrument Quality Criteria categories of "poor, fair, good, very good and excellent" developed by Fisher (2007). Citation: Fisher (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions 21*, 1095. Available from: http://www.rasch.org/rmt/rmt211.pdf

cogent education

5.3. Confirmation of dimensionality

Dimensionality was assessed by a Rasch principal components analysis (PCA) of the residuals within *Winsteps*. By completing a Rasch PCA of the residuals (referred to as the first contrast), evidence of a component that explains a large percentage of variance (usually >60%) of the residuals and a PCA eigenvalue for the first contrast of <3.0 are expected (Bond & Fox, 2015). For the PSDQ–S composite scale, the percentage of variance accounted for by the first factor was 69.2% with an eigenvalue of 9.6 (see Table 4). The percentage of unexplained variance in contrasts 1–5 of the PCA of the residuals was 1.4% (with a desired percentage of <5%) (see Table 4) for the PSDQ–S composite scale.

For the 11 PSDQ–S subscales, the percentage of variance accounted for by the first factor ranged from 52.1% to 87.8% with eigenvalues ranging from 1.6 to 2.3 (see Table 4). The percentage of unexplained variance in contrasts 1–5 of the PCA of the residuals ranged from 0.0% to 0.1% (see Table 4) for the 11 PSDQ–S subscales.

5.4. Item hierarchical ordering

The item hierarchical ordering of the PSDQ–S composite scale and 11 subscale items were examined. The item logit scores for the PSDQ–S composite scale ranged from −0.73 to 0.88 and for the 11 subscale items ranged from −0.93 to 1.41 (see Tables 2 and 3) with item means ranging from 0.0 logits to 0.0 (SD = 0.73) (see Table 4). The item separation index for the PSDQ–S composite scale was 5.71 and the number of separate item strata was 8 (see Table 4) while for the 11 subscales, the item separation index results ranged from 1.08 to 7.58 and the separate item strata numbers ranged from 1.8 to 10.1 (see Table 4).

The Wright Person-Item maps are located in Figures 1 and 2. They report the person ability distributions of the 117 participants mapped against the logit item difficulty scores for the PSDQ–S composition scale and 11 subscales. It provides a visual representation of the PSDQ–S hierarchical ordering of the items. In Figure 1, it appears that several of the PSDQ–S composite scale items had similar difficulty levels. For example, PSDQ–S composite scale items 5 and 7 had a logit score of −0.02, items 1, 17, and 24 had a logit score of −0.03, items 13, 15, and 18 had a logit score of −0.29, and items 28 and 37 had a logit score of −.53. From a hierarchical ordering perspective, there may be some item redundancy in relation to item difficulty for the PSDQ–S composite scale. However, given the purpose of the PSDQ–S is not competency or ability based, this is not necessarily relevant. For the 11 PSDQ–S subscales, no overlap in the hierarchical ordering of the items was observed.

5.5. Differential item functioning (DIF)

The PSDQ–S composite scale and 11 subscales were examined for DIF based on gender (males versus females). The PSDQ–S composite scale had one item that exhibited DIF based on gender, that being item 25. For the 11 subscales, only one subscale had one item that demonstrated DIF based on gender, that being item 29 on the Global Physical subscale (see Table 4). Overall, the PSDQ–S composite scale and 11 subscale items exhibited minimal DIF based on gender.

5.6. Person- and item-separation reliability

Person and item reliability indices were calculated. “Person separation reliability measures how accurately persons can be differentiated on the measured variable, whereas item separation reliability refers to how well the test distinguishes between items along the measured variable” (Teman, 2013, p. 420). These reliability coefficients can be interpreted in a similar manner as a Cronbach coefficient alpha (Boone et al., 2014). Person-separation reliability for the PSDQ–S composite scale was 0.90 and item separation-reliability was 0.97. For the 11 PSDQ–S subscales, the person-separation reliability ranged from 0.47 to 0.78 and item separation-reliability ranged from 0.61 to 0.98 (see Table 4). The Person Raw Score reliability (deemed equivalent to Cronbach’s alpha coefficient) for the PSDQ–S composite scale was 0.91 and for the 11 subscales ranged from 0.60 to 0.96 (see Table 4).

6. Discussion

The 40 PSDQ–S items were answered by a group of 117 typically developing school-age children ($\bar{x}$ age = 10 years, 2 months, SD = 1 year, 4 months). The RMM fit, dimensionality, hierarchical ordering of

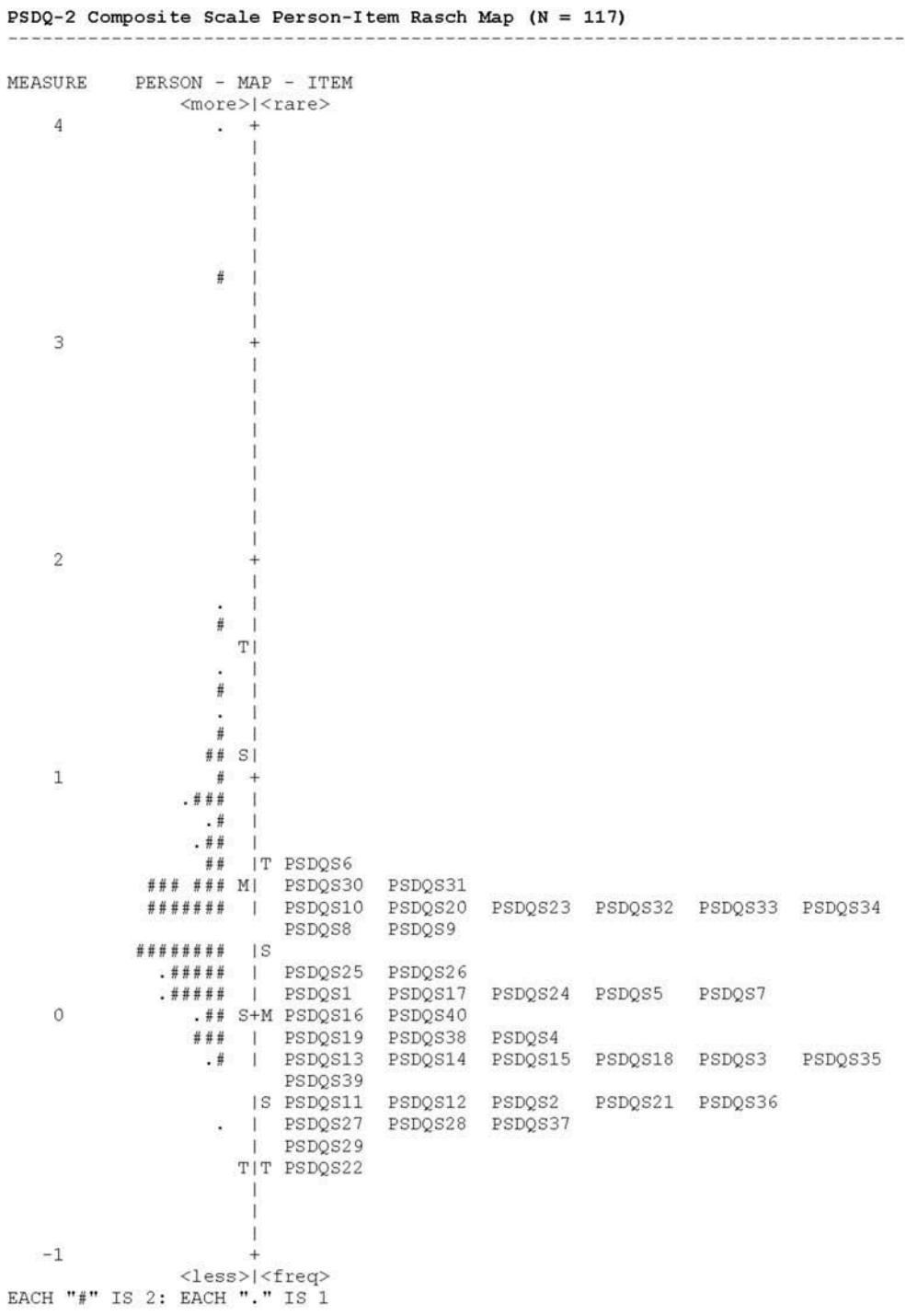

Figure 1. Wright person-item Rasch map for the PSDQ–S cpmposite scale (N = 117).

times, DIF based on gender, and person and item reliability were examined. Given that this is the first analysis of the PSDQ–S's structural validity using an item response-based approach, there is little to no comparison literature to contextualize the results of this study. Therefore, the findings of the RMM analysis of the PSDQ–S's composite scale and 11 subscales will be commented on in the context of the

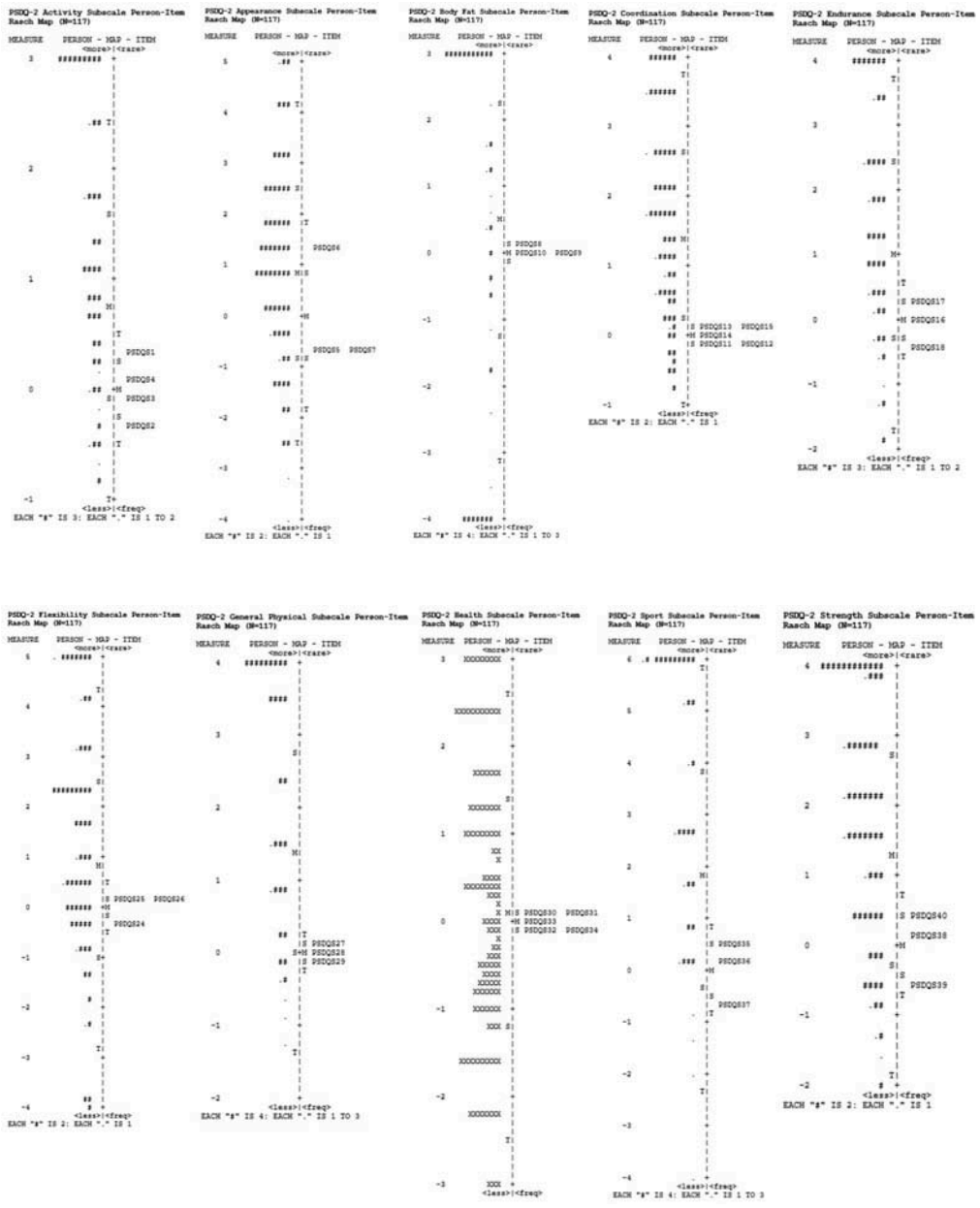

Figure 2. Wright person-item Rasch map for the PSDQ–2 subscales (N = 117).

general RMM expectations and what the implications of this are for the PSDQ–S. Comparison to peer-reviewed studies that utilized Classical Test Theory factor analytic approaches will be made where relevant.

6.1. RMM item fit

The PSDQ–S's composite scale and 11 subscale items all met the RMM Infit and Outfit MNSQ requirements, with the exception of one. For the PSDQ–S composite scale, one item misfit the RMM requirements, that being PSDQ–S Item 31. On the other hand, all the items for the 11 subscales were found to meet the RMM expectations. Up to now, the PSDQ-S has only been used to generate scores for the 11 subscales and not an overall "physical self-concept" composite score. However, the results of the RMM analysis indicate that the items of the PSDQ-S can be summed together to calculate a composite score. Overall, the PSDQ–S's composite scale and 11 subscales met the RMM fit requirements.

6.2. Dimensionality

Dimensionality of the PSDQ–S composite scale and its 11 subscales were examined via PCA of the residuals within *Winsteps*. The percentage of variance accounted for by the first factor was 69.2% with an eigenvalue of 9.6 for the PSDQ–S composite scale. The recommended values for the percentage of variance accounted for by the first factor is >60% and an eigenvalue of <2.0. Even though the eigenvalue is >2.0, the high percentage of variance accounted for indicates that the items that make up the PSDQ–S composite scale are unidimensional. It is possible that the dimensionality of the PSDQ–S composite scale could be improved by removing the one misfitting item, PSDQ–S Item 31. The RMM *Winsteps* analysis could then be rerun to determine if better model fit is attained and if the unidimensionality of the PSDQ–S composite scale improves.

For the 11 PSDQ–S subscales, the Health subscale had the highest percentage of variance accounted for by the first factor at 87.8% with an eigenvalue of 2.3, while the Global Physical subscale had the lowest percentage of variance explained by the first factor at 52.1% with an eigenvalue of 1.6. Only the Global Physical subscale had the percentage of explained variance lower than the desired 60% threshold which would therefore potentially indicate that this one PSDQ–S subscale is not unidimensional. In future, revisions of this subscale and the addition of one or two additional items might improve the breadth of the factor it represents. In summary, all but one of the 11 PSDQ–S subscales exhibited unidimensionality while the PSDQ–S composite scale met the RMM requirements for unidimensionality.

Marsh et al. (2010), the original authors of the PSDQ–S, used a Classical Test Theory-based cross-validation approach to provide evidence of the factor structure of the PSDQ-S. It was found that the factorial structure of the PSDQ–S subscales (nine distinct subscales and two global measures) was consistent across six different participant groups from three different cross-cultural contexts (e.g., Australia, Spain, and Israel). Given the fact that the PSDQ–S has been translated into Chinese, French, Finnish, and Slovenian (Dolenc, 2016; Haapea et al., 2016; Maïano et al., 2015; Wang et al., 2015), this indicates that it has demonstrated dimensional constancy when translated into other languages. While not directly analogous to the RMM dimensionality results, the factor structure findings reported by Marsh et al. (2010) and its factor structure consistency when translated into other languages, provides a point of comparison of the dimensionality of the 11 PSDQ–S subscales.

6.3. Hierarchical ordering

Referring to Figure 1, the Wright Person-Item map for the PSDQ–S composite scale charts the person ability logit scores against the item difficulty logit scores. It provides a visual representation of how the difficulty levels of the PSDQ–S composite scale match the ability levels of the participant group in a hierarchical representation. For the most part, there is a reasonable spread in the difficulty levels of the PSDQ–S composite scale items with item logit scores ranging from −0.73 to 0.88. However, there appears to be a mismatch between the person ability logit scores and the item difficulty logit scores in that there are a sizeable number of participants who appear to have higher ability levels than PSDQ–S composite scale item difficulty levels.

In other words, there is a lack of spread of item difficulty levels in the PSDQ–S composite scale to cover the full range of participant ability levels. There appear to be several items with the same level of difficulty. For example, on Figure 1, PSDQ–S composite scale items 10, 20, 23, 32, 33, and 34 appear to have the same level of item difficulty. There appears to be a similar case for PSDQ–S composite scale items 3, 13, 14, 15, 18, and 35.

Therefore, in future revisions of the PSDQ–S composite scale, it might be worth exploring the impact of discarding a few items that appear to have the same difficulty level and then generate some additional items that have higher levels of difficulty. However, this may be a moot point since the PSDQ–S composite scale is designed to measure a quantity or amount of the factor it claims to represent. The PSDQ–S was not designed to be an evaluative scale of a person's ability or skill. However, it is still worth

consideration of exploration since reducing the number of items with duplicate difficulty levels would in the end create a shorter, more targeted scale with less respondent burden.

In relation to the 11 PSDQ–S subscales, there is a limited range of item logit difficulty levels in comparison to the person ability logit results for all the scales. Likewise, there is a much broader range of person ability logit range scores. The average item difficulty is typically expected to be comparable to the average person ability (Bond & Fox, 2015). Again, this situation of a disparity between the person ability and item difficulty hierarchical ordering creates the situation of a mismatch or poor targeting between the difficulty levels of the PSDQ–S subscale items and the participants who answered the items (McHorney & Tarlov, 1995).

The four items that make up the PSDQ–S Activity subscale had a moderate range of difficulty levels (−0.28 to 0.35 logits), whereas the three items that make up the PSDQ–S Appearance subscale only had two levels of item difficulty (−0.70 to 1.41). The three items that make up the PSDQ–S Body Fat subscale have a very narrow item logit difficulty range (−0.6 to 0.12).

The five items included in the PSDQ–S Coordination subscale have a very limited item difficulty logit score range (−0.10 to 0.11) compared to the three items that make up the Endurance subscale (−0.46 to 0.35). The Global Esteem and Sport subscales have moderate item logit difficulty ranges, whereas the Flexibility, General Physical, and Health subscales have limited item logit difficulty ranges. Similar to the PSDQ–S composite scale, all 11 subscales exhibit an incongruity between the ability levels of the participants and the difficulty levels of the items. Given that the number of individual items per PSDQ–S subscale ranges between three and five and the fact that the subscales are not designed to measure a person's skill level or ability but instead provide a score in relation to the amount of a factor or variable, the misalignment between the person logit scores and the item logit scores may not be such a major concern. During the next revision of the PSDQ–S composite scale and subscales, however, the inclusion of additional items with a greater range of item difficulty is recommended. The provision of Wright Person-Item maps offers a window to the hierarchical alignment match or mismatch between scale item difficulty levels and the ability levels of the respondents who answer the items.

6.4. Differential Item Functioning (DIF)

DIF provides information about whether or not the items of a scale are biased or act differently when completed by a specific subgroup of participants (Bond & Fox, 2015). This is particularly important when a scale or instrument is high-stakes or would have potential consequences for the test-takers. It also provides valuable information about specific items that may need to be revised or ultimately discarded. The PSDQ–S composite scale and its 11 subscales were scrutinized for DIF based on gender (male versus female participants).

PSDQ–S composite scale item 25 and Global Physical subscale item 29 demonstrated DIF based on gender. In short, the items of the 11 subscales and composite scale revealed minimal DIF based on gender. Given the PSDQ–S has been translated into Chinese, French, Finnish, and Slovenian (Dolenc, 2016; Haapea et al., 2016; Maïano et al., 2015; Wang et al., 2015), further examination of the DIF of the PSDQ–S in relation to the first language of the respondents is recommended.

6.5. Person and item reliability

Person reliability results ranged from 0.47 to 0.78 for the 11 PSDQ–S subscales, which was less than optimal. The desired level is to have person reliability scores of >0.80. The low person reliability scores can be partially explained by the number of items per PSDQ–S subscale (which ranged from a minimum of three items to a maximum of five items) (Marsh et al., 2010). The misalignment between the person ability logit scores and the item difficulty logit scores for all 11 PSDQ–S subscales could also be an explanation for the low person reliability scores.

The person separation index (PSI) for the 11 subscales ranged from 0.95 to 1.86. The recommended level for the PSI is >2.0. The PSI for the PSDQ–S composite scale was 1.16.

The Person Raw Score Reliability (deemed comparable to Cronbach's coefficient alpha) for the 11 subscales ranged from 0.72 to 0.96, and 0.60 was reported for the composite scale. The recommended level for the Person Raw Score Reliability is >0.80. In short, the person reliability indices for the PSDQ–S composite scale and 11 subscales were in the poor to excellent ranges.

The item reliability coefficients for the 11 PSDQ–S subscales ranged from 0.61 to 0.98. For the PSDQ–S composite scale the item reliability coefficient was 0.97. The desired range for the item reliability coefficient is >0.80. The Global Physical subscale had the lowest item reliability coefficient.

7. Limitations

Limitations of this study include the convenience sampling approach used to recruit participants. Also, participants were recruited from one geographical region which may be a source of sampling bias. Since this study did not have any formal funding, it was not feasible or possible to recruit participants from a wider area. Only children who were typically developing were included in the sample.

8. Conclusion

The PSDQ–S is a useful scale that provides a multi-faceted overview of participants' views of their physical self. It is appropriate for use in clinical, education, and research settings. It has also been translated into several other languages which demonstrates its usefulness and relevance in cross-cultural contexts. As mentioned in the *Standards* (AERA, APA, & NCME, 2014), the validity of a scale is ongoing, dynamic, and ever-evolving. Contributing to a scale's validity evidence is also a group effort from a dedicated community of scholars, scale users, and scale consumers. Using the RMM, aspects of the PSDQ–S's structural validity in relation to its composite scale and 11 subscales were examined.

The findings indicated that the dimensionality of the PSDQ–S composite scale and 11 subscales was supported based on the RMM Infit and ZSTD scores. Only one PSDQ–S composite scale item demonstrated RMM misfit. This provides evidence of the structural validity of the PSDQ–S composite scale and its 11 subscales. The PSDQ–S's composite scale and 11 subscales also exhibited minimal DIF based on gender. Only one item in the PSDQ–S's composite scale and one item on the Global Physical subscale exhibited DIF. Likewise, the RMM Wright Person-Item maps also provided insights into the PSDQ–S composite scale and 11 subscales' hierarchical item ordering. What's more, the PSDQ–S composite scale and 11 subscales also exhibited reasonable to moderate levels of person and item reliability. Overall, the structural validity of the PSDQ–S composite scale and 11 subscales was supported based on the RMM analysis findings.

In conclusion, it is recommended this study be replicated with a larger sample size recruited from a wider geographical region and with participants who have been randomly selected. It is also recommended that other aspects of the PSDQ–S's validity be examined such as its relationship to other variables (convergent and divergent validity), consequences to participants of testing, and criterion validity. Further investigation of the DIF of the PSDQ–S items in relation to other participant traits (such as education background, age, and ethnicity) is suggested. Likewise, further examination of the PSDQ–S's test-retest reliability and cross-cultural validity is warranted.

Funding
The study received no funding from any source.

Ethical approval
Ethical approval was granted by Monash University Human Research Ethics Committee approval on 9 April 2015 (Ethics approval ID: 2015-6069-5898).

Authors' contributions
Dr. Ted Brown designed the study, collected the data, performed the statistical analyses, and completed the interpretation of the data findings. Both authors drafted the manuscript as well as read and approved the final version of the manuscript.

Author details
Ted Brown[1]
E-mail: education@cogentoa.com
ORCID ID: http://orcid.org/0000-0001-9403-5877
Tore Bonsaksen[2,3]
E-mail: tore.bonsaksen@oslomet.no
ORCID ID: http://orcid.org/0000-0001-6315-1111
[1] Department of Occupational Therapy, School of Primary and Allied Health Care, Faculty of Medicine, Nursing and Health Sciences, Monash University, Peninsula Campus, Frankston 3199, Australia.
[2] Department of Occupational Therapy, Prosthetics and Orthotics, Faculty of Health Sciences, OsloMet - Oslo Metropolitan University, Oslo, Norway.

[3] Faculty of Health Studies, VID Specialized University, Sandnes, Norway.

Citation information
Cite this article as: An examination of the structural validity of the Physical Self-Description Questionnaire-Short Form (PSDQ–S) using the Rasch Measurement Model, Ted Brown & Tore Bonsaksen, *Cogent Education* (2019), 6: 1571146.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arnadottir, G., & Fisher, A. (2008). Rasch analysis of the ADL scale of the A-ONE. *American Journal of Occupational Therapy*, *62*, 51–60. doi:10.5014/ajot.62.1.51

Babic, M. J., Morgan, P. J., Plotnikoff, R. C., Lonsdale, C., White, R. L., & Lubans, D. R. (2014). Physical activity and physical self-concept in youth: Systematic review and meta-analysis. *Sports Medicine*, 44, 1589–1601. doi:10.1007/s40279-014-0229-z

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge/Taylor and Francis Group.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht, Netherlands: Springer.

Brewer, W. A., & Olson, S. L. (2015). Are there relationships between perceived and actual measures of physical fitness and health for healthy young women? *Comprehensive Psychology*, 4, 2. Retrieved from http://journals.sagepub.com/doi/pdf/10.2466/06.CP.4.2

Brown, T. (2010). Construct validity: A unitary concept for occupational therapy assessment and measurement. *Hong Kong Journal of Occupational Therapy*, *20*(1), 30–42. doi:10.1016/S1569-1861(10)70056-5

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 212–225. doi:10.1080/0969594X.2015.1063479

Cremeens, J., Eiser, C., & Blades, M. (2006). Characteristics of health-related self-report measures for children aged three to eight years: A review of the literature. *Quality of Life Research*, *15*, 739–754. doi:10.1007/s11136-005-4184-x

Crocker, P. R. E., Sabiston, C. M., Kowalski, K. C., McDonough, M. H., & Kowalski, N. (2006). Longitudinal assessment of the relationship between physical self-concept and health-related behavior and emotion in adolescent girls. *Journal of Applied Sports Psychology*, *18*(3), 185–200. doi:10.1080/10413200600830257

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

Dishman, R. K., Hales, D. P., Almeida, M. J., Pfeiffer, K. A., Dowda, M., & Pate, R. R. (2006). Factorial validity and invariance of the physical self-description questionnaire among black and white adolescent girls. *Ethnicity & Disease*, *16*(2), 551–558.

Dolenc, P. (2016). The short form of the physical self-description questionnaire: Validation study among Slovenian elementary and high school students. *Journal of Psychological and Educational Research*, *24* (2), 58–74. doi:10.1016/j.bodyim.2014

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37* (9), 830–837. doi:10.1046/j.1365-2923.2003.01594.x

Ekeland, E., Heian, F., & Hagen, K. B. (2005). Can exercise improve self-esteem in children and young people: A systematic review of randomized controlled trials. *British Journal Sports Medicine*, *39*(11), 792–798. doi:10.1136/bjsm.2004.017707

Fisher, W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*, 1095. Retrieved from http://www.rasch.org/rmt/rmt211.pdf

Goodwin, L. D. (2002). Changing conceptions of measurement validity: An update on the new standards. *Journal of Nursing Education*, *41*(3), 100–106. doi:10.3928/0148-4834-20020301-05

Haapea, I., Haverinen, K., Honkalampi, K., Kuittinen, M., & Räty, H. (2016). The factor structure and reliability of the short form of the physical self-description questionnaire in a finnish adolescent athlete sample. *International Journal of Sport and Exercise Psychology*, published online. doi:10.1080/1612197X.2016.1266504

Klomsten, A. T., Skaalvik, E. M., & Espnes, G. A. (2004). Physical self-concept and sports: Do gender differences still exist? *Sex Roles*, *50*(1–2), 119–127. doi:10.1023/B:SERS.0000011077.10040.9a

Lim, M., Rodger, S., & Brown, T. (2009). Using Rasch analysis for establishing the construct validity of rehabilitation assessment tools. *International Journal of Therapy and Rehabilitation*, *16*(5), 251–260. doi:10.12968/ijtr.2009.16.5.42102

Linacre, J. M. (2011). *Winsteps (version 3.70.0)* [Computer software]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2013). Misfit diagnosis: Infit outfit mean-square standardized. Retrieved from: http://www.winsteps.com/winman/index.htm?diagnosingmisfit.htm

Linacre, J. M., & Wright, B. D. (1998). *A user's guide to bigsteps winsteps Rasch-Model computer program*. Chicago, IL: MESA Press.

Maïano, C., Morin, A. J., & Mascret, N. (2015). Psychometric properties of the short form of the physical self-description questionnaire in a French adolescent sample. *Body Image*, *12*, 89–97. doi:10.1016/j.bodyim.2014.10.005

Marsh, H. (1999). *Physical self-description questionnaire (complete) package*. Oxford, UK: SELF Research Centre.

Marsh, H. W. (1996a). Construct validity of physical self-description questionnaire responses: Relations to external criteria. *Journal of Sport and Exercise Psychology*, *18*, 111–131. doi:10.1123/jsep.18.2.111

Marsh, H. W. (1996b). Physical self-description questionnaire: Stability and discriminant validity. *Research Quarterly for Exercise and Sport*, *67*(3), 249–264. doi:10.1080/02701367.1996.10607952

Marsh, H. W. (1997). The measurement of physical self-concept: A construct validation approach. In K. Fox (Ed.), *The physical self-concept: From motivation to well-being* (pp. 27–58). Champaign, IL: Human Kinetics.

Marsh, H. W., Martin, A. J., & Jackson, S. (2010). Introducing a short version of the physical self-description questionnaire: New strategies, short-form evaluative criteria, and applications of factor analyses. *Journal of Sport and Exercise Psychology*, *32*(4), 438–482. doi:10.1123/jsep.32.4.438

Marsh, H. W., Richards, G. E., Johnson, S., Roche, L., & Tremayne, P. (1994). Physical self-description questionnaire: Psychometric properties and a multitrait-multimethod analysis of relations to existing instruments. *Journal of Sport and Exercise Psychology*, *16*, 270–305. doi:10.1123/jsep.16.3.270

Marsh, H. W., Tomás Marco, I., & Hülya Aþçý, F. (2002). Cross-cultural validity of the physical self-description questionnaire: Comparison of factor structures in Australia, Spain and Turkey. *Research Quarterly for Exercise and Sport*, *73*(3), 257–270. doi:10.1080/02701367.2002.10609019

Martin, J. J., & Whalen, L. (2012). Self-concept and physical activity in athletes with physical disabilities. *Disability and Health Journal*, *5*(3), 197–200. doi:10.1016/j.dhjo.2012.03.006

McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: Are available health status surveys adequate. *Quality of Life Research*, *4*, 293–307. doi:10.1007/BF01593882

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–104). New York, NY: American Council on Education and Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. doi:10.1037/0003-066X.50.9.741

Mokkink, L., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., … de Vet, H. C. W. (2006). Protocol of the COSMIN study: Consensus-based standards for the selection of health measurement instruments. *BMC Medical Research Methodololgy*, *6*(1), 2. doi:10.1186/1471-2288-6-2

Mokkink, L., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., … de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737–745. doi:10.1016/j.jclinepi.2010.02.006

Newton, P. E., & Baird, J.-A. (2016). The great validity debate. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 173–177. doi:10.1080/0969594X.2016.1172871

Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 178–197. doi:10.1080/0969594X.2015.1037241

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*(1), 108–116. doi:10.7334/psicothema2013.260

Simons, J., Capio, C. M., Adriaenssens, P., Delbroek, H., & Vandenbussche, I. (2012). Self-concept and physical self-concept in psychiatric children and adolescents. *Research in Developmental Disabilities*, *33*(3), 874–881. doi:10.1016/j.ridd.2011.12.012

Smith, R. M. (1992). *Applications of Rasch measurement*. Chicago, IL: MESA Press.

Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In W. M. Greenwich (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 316–328). Norword, NJ: Ablex.

Streiner, D., Norman, G., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). Oxford, UK: Oxford University Press.

Teman, E. D. (2013). A Rasch analysis of the statistical anxiety rating scale. *Journal of Applied Measurement*, *14*(4), 414–434.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Ulrika, A., Johan, P., & Guy, M. (2017). Correspondence between physical self-concept and participation in, and fitness change after, biweekly body conditioning classes in sedentary women. *The Journal of Strength and Conditioning Research*, *31*(2), 451–461. doi:10.1519/JSC.0000000000001721

Utesch, T., Dreiskämper, D., Naul, R., & Geukes, K. (2018). Understanding physical (in-) activity, overweight, and obesity in childhood: Effects of congruence between physical self-concept and motor competence. *Scientific Reports*, *8*, 5908. doi:10.1038/s41598-018-24139-y

Wang, C. K. J., Sun, Y., Liu, W. C., Yao, J., & Pyun, D. Y. (2015). Latent profile analysis of the physical self-description among Chinese adolescents. *Current Psychology*, *34*(2), 282–293. doi:10.1007/s12144-014-9257-y

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94. doi:10.20982/tqmp.09.2.p079

Cogent Education **(ISSN: 2331-186X) is published by Cogent OA, part of Taylor & Francis Group.**